AgriEngineering

MDPI

Article

Caffeine Content Prediction in Coffee Beans Using Hyperspectral Reflectance and Machine Learning

Dthenifer Cordeiro Santana [1], Rafael Felipe Ratke [1], Fabio Luiz Zanatta [2], Cid Naudi Silva Campos [1], Ana Carina da Silva Cândido Seron [1], Larissa Pereira Ribeiro Teodoro [1], Natielly Pereira da Silva [1], Gabriela Souza Oliveira [1], Regimar Garcia dos Santos [3], Rita de Cássia Félix Alvarez [1], Carlos Antonio da Silva Junior [4], Matildes Blanco [1] and Paulo Eduardo Teodoro [1,*]

[1] Agronomy Department, Federal University of Mato Grosso do Sul (UFMS), Chapadão do Sul 79560-000, MS, Brazil; dthenifer.santana@ufms.br (D.C.S.); rafael.ratke@ufms.br (R.F.R.); cid.campos@ufms.br (C.N.S.C.); ana.candido@ufms.br (A.C.d.S.C.S.); larissa_ribeiro@ufms.br (L.P.R.T.); natielly.silva@ufms.br (N.P.d.S.); gabriela_souza@ufms.br (G.S.O.); rita.alvarez@ufms.br (R.d.C.F.A.); m-blanco@ufms.br (M.B.)

[2] Agronomy Department, Professor Cinobelina Elvas Campus, Federal University of Piauí, Bom Jesus 58930-000, PI, Brazil; fabio.zanatta@ufpi.edu.br

[3] Plant Sciences Building, Department of Horticulture, The University of Georgia, Athens, GA 30602, USA; regimar.garcia@uga.edu

[4] Department of Geography, State University of Mato Grosso (UNEMAT), Sinop 78555-000, MT, Brazil; carlosjr@unemat.br

* Correspondence: eduteodoro@hotmail.com

Abstract: The application of hyperspectral data in machine learning models can contribute to the rapid and accurate determination of caffeine content in coffee beans. This study aimed to identify the machine learning algorithm with the best performance for predicting caffeine content and to find input data for these models that can improve the accuracy of these algorithms. The coffee beans were harvested one year after the seedlings were planted. The fresh beans were taken to the spectroscopy laboratory (Laspec) at the Federal University of Mato Grosso do Sul, Chapadão do Sul campus, for spectral evaluation using a spectroradiometer. For the analysis, the dried coffee beans were ground and sieved for the quantification of caffeine, which was carried out using a liquid chromatograph on the Waters Acquity 1100 series UPLC system, with an automatic sample injector. The spectral data of the beans, as well as the spectral data of the roasted and ground coffee, were analyzed using machine learning (ML) algorithms to predict caffeine content. Four databases were used as input: the spectral information of the bean (CG), the spectral information of the bean with additional clone information (CG+C), the spectral information of the bean after roasting and grinding (CGRG) and the spectral information of the bean after roasting and grinding with additional clone information (CGRG+C). The caffeine content was used as an output to be predicted. Each database was subjected to different machine learning models: artificial neural networks (ANNs), decision tree (DT), linear regression (LR), M5P, and random forest (RF) algorithms. Pearson's correlation coefficient, mean absolute error, and root mean square error were tested as model accuracy metrics. The support vector machine algorithm showed the best accuracy in predicting caffeine content when using hyperspectral data from roasted and ground coffee beans. This performance was significantly improved when clone information was included, allowing for an even more accurate analysis.

Keywords: support vector machine; spectroscopy; secondary metabolites

Citation: Santana, D.C.; Ratke, R.F.; Zanatta, F.L.; Campos, C.N.S.; Seron, A.C.d.S.C.; Teodoro, L.P.R.; Silva, N.P.d.; Oliveira, G.S.; Santos, R.G.d.; Alvarez, R.d.C.F.; et al. Caffeine Content Prediction in Coffee Beans Using Hyperspectral Reflectance and Machine Learning. *AgriEngineering* **2024**, *6*, 4480–4492. https://doi.org/10.3390/agriengineering6040255

Academic Editors: Sotirios K. Goudos, Achilles Boursianis and Shaohua Wan

Received: 1 November 2024
Revised: 15 November 2024
Accepted: 22 November 2024
Published: 26 November 2024

1. Introduction

Coffee is one of the most popular beverages worldwide due to its high commercial value and flavor, which is influenced by its chemical composition and the treatment the grains receive from the moment they are harvested until they are processed [1]. It is a

commodity with high economic value in the world, second only to oil in terms of commercialized value; coffee cultivation is a major contributor to the socioeconomic development of developing tropical countries, with Brazil being among the world's largest producers of the bean [2]. In this way, coffee is a traditional and widely consumed drink in Brazil, being crucial in the global and Brazilian economy [3].

Among the variables used to assess coffee quality are caffeine, trigonelline and chlorogenic acids [4]. Caffeine is said to be one of the components of coffee beans that is responsible for the bitterness and stimulating effect of the drink that makes it so well known [5]. In addition, caffeine is considered an important indicator of coffee quality, and its quantification is commonly established by chemical methods, which tend to be time-consuming, destructive and expensive, making a rapid determination of this content unfeasible [6].

Caffeine in plants is a type of secondary metabolite belonging to the alkaloid class. These low molecular weight compounds act as bioactive signaling molecules, providing protection against stressors. As a result, they contribute to strengthening the plant's defenses, in addition to promoting vegetative and/or reproductive growth, thus improving yield [7] and acting as a natural defense, helping the plant to protect itself against herbivores and pests, due to its toxic properties in high concentrations. The accumulation and concentration of caffeine are influenced by genetic and environmental factors, occurring predominantly in leaves, seeds and fruits of plants, such as the coffee plant [3].

One of the most widely used methods for determining caffeine is through liquid chromatography, which is used to analyze coffee beans and beverages, following protocols that have high sensitivity, precision and accuracy. However, it is a highly complex methodology that requires lengthy and expensive sample preparation (cleaning) [1,8–10]. This creates an opportunity for proposing new, faster methods of caffeine content determination. The use of hyperspectral information is an innovative approach in the food sector that enables rapid analysis of food materials [1].

Obtaining hyperspectral information can provide several advantages, such as quickly and conveniently collecting spectral information from a large number of coffee samples in a short period [11]. Hyperspectral data combine the non-destructive and fast nature of NIRS with highly detailed information on the heterogeneity of the sample being studied [12]. Therefore, the use of sensors that provide hyperspectral information is a valuable tool for predicting caffeine in coffee beans, helping to improve both the efficiency of the process and the quality of the final product. However, the amount of data generated by these sensors is extensive, making it difficult to use traditional data analysis for such a prediction.

The use of machine learning (ML) can efficiently overcome the problem of the large volume of data generated by the sensor and the lack of direct correlation between the spectral data of the coffee beans and the caffeine content. ML techniques are a data analysis process used to make complex decisions. It represents a branch of artificial intelligence (AI) that allows computers to "learn" from the available data [13]. The use of spectral data in conjunction with machine learning techniques to predict caffeine content in plants, such as coffee, is an innovative and efficient approach that is not well researched. The integration of these techniques has not yet been proposed in an efficient way and using different ML models with different input configurations seeking the best performances. Additional input information to the spectral variables, such as information from the different genotypes under analysis, should be tested, since the accuracy of caffeine prediction can be variable depending on the genetic material information. This is because there may be differences in caffeine content between genetic materials, and this behavior can be captured by machine learning models and have an impact on the models' ability to learn and generalize.

In light of the above, the application of hyperspectral data in machine learning models can contribute to the rapid and accurate determination of the caffeine content in coffee beans. This study aimed to identify the machine learning algorithm with the best performance for predicting caffeine content and to find input data for these models that can improve the accuracy of these algorithms.

2. Materials and Methods

2.1. Experiment

The experiment was carried out at Laranjeiras' Farm, in the city of Currais-PI. Samples of beans from 15 coffee cultivars (Conilon and Arabica) were evaluated in a randomized block design with four replications. Coffee was grown using drip irrigation. The soil at the site of the experiment was characterized as physiochemical through analysis of the soil sampled in the experiment (Table 1). The soil was characterized as a typical dystrophic yellow latosol (SIBCS 3° edition).

Table 1. Chemical and granulometric characterization of the experiment area.

Depth	pH	Ca	Mg	Al	H + Al	K	SB	T	P	V	M	MO	Sand	Silt	Clay
m	$H\ O_2$	----------$cmol_c\ dm^3$ ----------							$mg\ dm^3$	-----%-------			----------g kg----------		
0–0.20	6.1	3.51	0.55	0.00	2.31	0.25	4.31	6.62	24.91	65.1	0.0	15.3	851	5	144
0.20–0.40	4.9	0.56	0.15	0.10	1.82	0.05	0.76	2.57	4.23	29.5	11.6	4.0	823	9	167
0.40–0.60	4.9	0.37	0.13	0.10	4.62	0.05	0.54	5.16	1.90	10.5	15.6	1.4	770	18	213

pH = hydrogen potential; P = phosphorus; K = potassium; Ca = calcium; Mg = magnesium; Al = aluminum; H + Al = hydrogen + aluminum; SB = sum of bases; T = cation exchange capacity at pH 7.0; m = aluminum saturation; V = base saturation; MO = organic matter.

2.2. Spectral Analysis

The coffee beans were harvested one year after the seedlings were planted. The fresh beans were taken to the spectroscopy laboratory (Laspec) at the Federal University of Mato Grosso do Sul, Chapadão do Sul campus, for spectral evaluation using a spectroradiometer (FieldSpec 4 HRes, Analytical Spectral Devices, Boulder, CO, USA), which provides spectral information in the 350 to 2500 nm range. The spectral data were used to form the spectral signature of the coffee beans for each clone evaluated (Figure 1). The hyperspectral data were preprocessed using a Savitzky–Golay (SG) smoothing filter and mean normalization, as suggested by [14]. Each coffee sample was subdivided into six samples for spectral analysis.

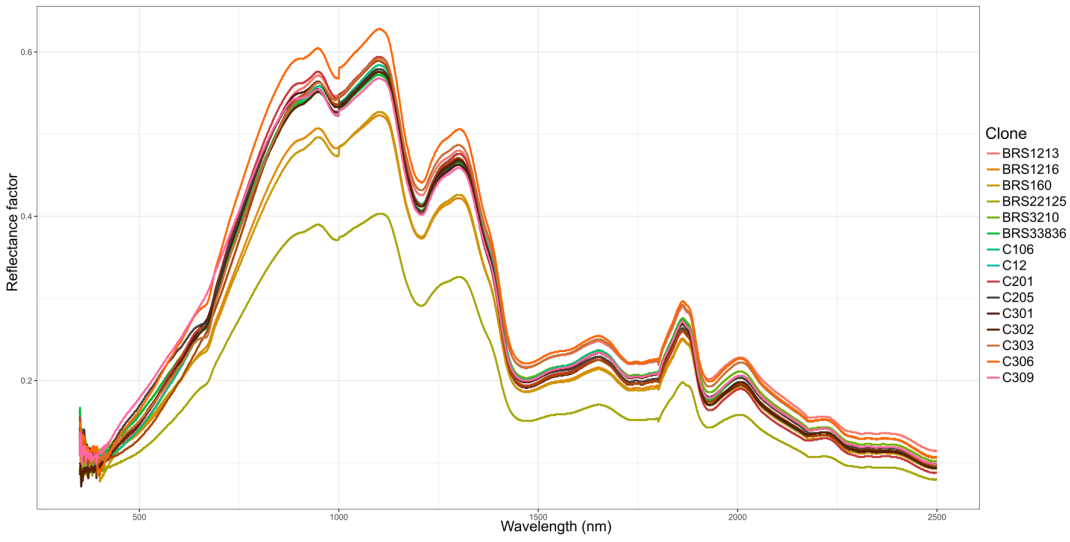

Figure 1. Hyperspectral curve of coffee beans from each clone evaluated.

After the spectral evaluation of the grains, they were sent for milling and medium roasting, following strict recommendations for this procedure. The roasts were carried out using the reference parameters of the SCA (Specialty Coffee Association) methodology.

According to the methodology, roasting must be standardized to guarantee the correct sensory evaluation of the batches, without penalizing or rewarding a batch due to the

roasting standard. The samples were roasted in the Carmomaq Laboratto sample roaster designed to analyze specialty coffees. The roaster uses gas heating, with precise control of gas flow and airflow, allowing independent adjustments. This level of control is necessary to ensure a consistent roasting pattern. After the roasting rest, the samples were subjected to grinding. We used the Mahlkonig EK43 grinder, known for its precision and consistency, guaranteeing a uniform grind suitable for tasting. The samples were then returned to Laspec for a second hyperspectral evaluation of the grains after they had been roasted and ground (Figure 2).

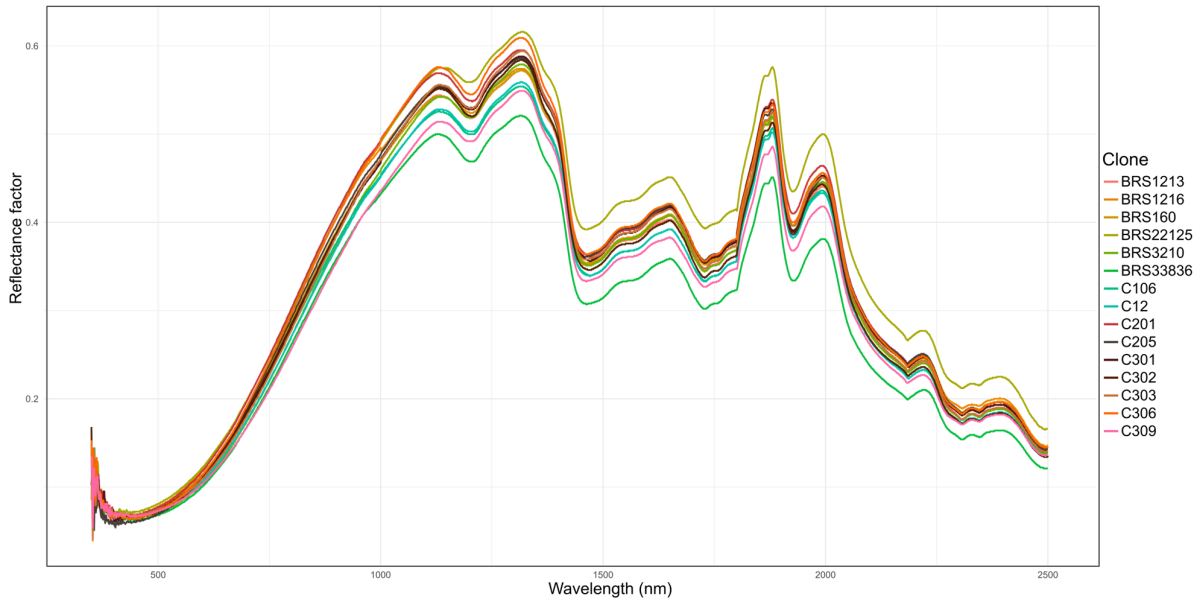

Figure 2. Hyperspectral curve of ground and roasted coffee beans from each clone evaluated.

2.3. Determination of Caffeine

For the analysis, the dried coffee beans were ground and sieved. An aliquot of 0.05 g of the samples was added to test tubes, after which 0.2 g of magnesium oxide was added. A total of 10 mL of Milli-Q water was added to the test tubes, which were then ultrasonicated for 5 min. The tubes were then placed in a water bath at 100 °C for 20 min and homogenized every 5 min. After this period, Milli-Q water was added to complete the total 20 mL volume. Then, 1.0 mL of each sample was filtered through a syringe with a 0.2 µm filter and transferred to 1.5 mL vials before injection into an ultra-performance liquid chromatography (UPLC) system. Aliquots of 1.0 µL were used for direct injection into the equipment. Each sample was analyzed three times.

Caffeine content was quantified using a Waters Acquity 1100 series UPLC liquid chromatograph with an automatic sample injector. The analyses were carried out using a 1.7 µm BEH C18 reverse phase column (internal diameter 2.1 mm (i.d.) £ 50 mm). A binary linear gradient system was used, with Milli-Q water as solvent A and acetonitrile as solvent B as mobile phases. The initial gradient was 99% for solvent A and 1.0% for solvent B from 0 to 1 min, 50% A and 50% B from 1 to 1.01 min, 5% A and 95% B from 1.01 to 1.10 min, returning to 99% A and 1% B at 1.10 min and remaining like this until 3 min, which was the run time for each sample. The mobile phase flow rate was 0.4 mL min^{-1} and the column temperature during the run was 40 °C.

Caffeine was detected using a Waters photodiode array detector, set to a wavelength of 254 nm. Commercially purchased standards were used to detect caffeine at the following concentrations: 0.002; 0.01; 0.03; 0.04; 0.05; 0.08; 0.1;0.16 and 4.0 mg mL^{-1}. The qualitative and quantitative identity of the peak was confirmed by comparing the retention times and UV spectra of the individual compounds using the standard addition method.

All the solvents used in the chromatographic analysis were HPLC grade, and before use, they were vacuum filtered through a 0.2 µm pore membrane and then degassed in a vacuum system using ultrasound. The water used was distilled and then ultra-purified in a Milli-Q system before being degassed.

2.4. Machine Learning Analysis

The spectral data of the beans and the spectral data of the roasted and ground coffee were analyzed using machine learning (ML) algorithms for caffeine content prediction. A flowchart of analyses performed is shown in Figure 3. A total of 108 samples were used in each dataset. Four databases were used as input: the spectral information of the bean (CG), the spectral information of the bean with additional clone information (CG+C), the spectral information of the bean after roasting and grinding (CGRG) and the spectral information of the bean after roasting and grinding with additional clone information (CGRG+C). The caffeine content was used as an output to be predicted. After identifying the best input for predicting caffeine content, it was resubmitted to the ML models with different sample sizes: all the samples evaluated (ALL), i.e., consisting of the 108 samples; half the samples evaluated (50%); and two-thirds of the samples evaluated (75%). Samples used to compose the partitioned datasets (50 and 75%) were randomly chosen from the original dataset (ALL).

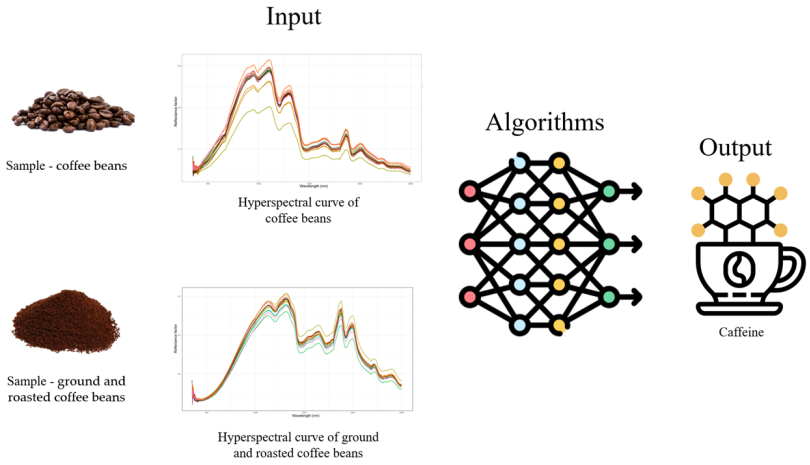

Figure 3. Flowchart of analyses performed.

Each database was subjected to different machine learning models, artificial neural networks (ANNs), decision tree (DT), linear regression (LR), M5P algorithm, random forest (RF) and support vector machine (SVM), which were selected according to [15–17]. The software used for the analyses was Weka 3.9.4, with cross-validation stratified 10 times with 10 repetitions using the standard (default) configuration for all the models tested, except for ANN, which will use two layers with ten neurons in each as suggested in the methodology of [18]. Default hyperparameter configuration consists of a DT algorithm (named REPTree), which builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning (with backfitting). M5P is a regression M5′ model tree algorithm, whose default hyperparameters carry out pruning and a minimum number of instances to allow at a leaf node equal to 4.0. ANN tested is the Multilayer Perceptron type, whose default configuration considers a learning rate of 0.3, momentum equal to 0.2 and number of epochs equal to 500. Finally, the default configuration for the RF considers no restriction on the size of the trees and the number of trees equal to 100, while SVM consists of a regression learner (called SMOreg) whose kernel type is polynomial, with the parameter C (cost) equal to 1.0, epsilon parameter equalt to 0.001 and tolerance parameter equal to 0.001.

Pearson's correlation coefficient (r), mean absolute error (MAE) and root mean square error (RMSE) were tested as metrics of model accuracy. To check the significance of the inputs, the algorithms tested and the interaction between both, an analysis of variance was carried out using a completely randomized design in a factorial scheme. If statistically significant, bar graphs were generated with the means of r, MAE and RMSE, grouped according to the Scott–Knott test at 5% probability. The grouping of means and all the graphs in the manuscript were generated using the ggplot2 and ExpDes.pt packages in the R software 4.1.0 version. A second ML analysis was then carried out with different sample sizes using the best input found in the first phase of the ML analysis.

3. Results

3.1. Basic Information on Caffeine

The caffeine content of each genetic material varied, with BRS1213, BRS22125, C301, C303 and C309 showing the highest amounts (Figure 4). The lowest amount of caffeine was observed in clone 106. The mean comparison test thus reaffirms the variability of caffeine among the clones studied. In addition to the variability among clones, there is also a variation in caffeine content within the clones themselves.

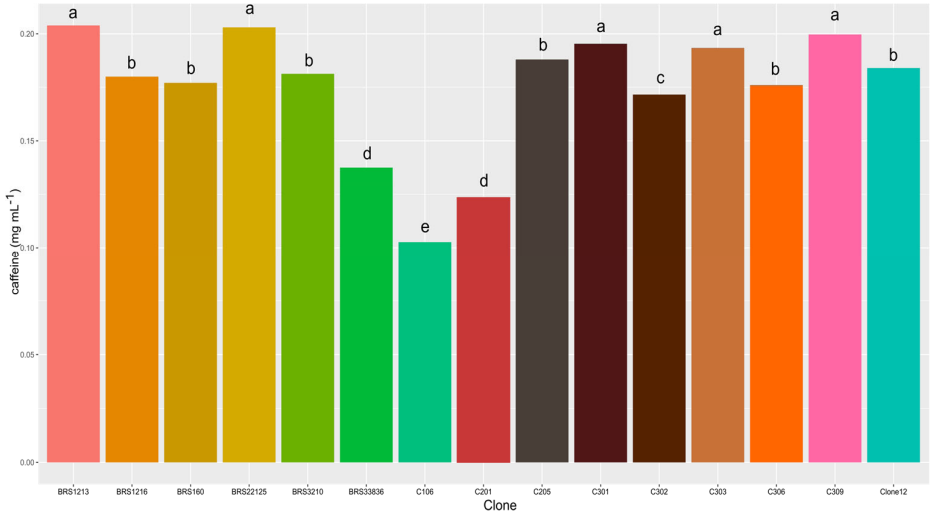

Figure 4. Comparison of means for caffeine content in the beans of different coffee clones by grouping means using the Scott-Knott test.

3.2. Caffeine Prediction Using Machine Learning

Since there was variability in the caffeine content of the coffee samples tested, the spectral data were applied to ML models to predict the caffeine content of the samples. The ML and input sources of variation were significant, as was the interaction among them (Table 2), for all accuracy tests to which the models were subjected.

Table 2. Mean squares of the analysis of variance for the Pearson correlation coefficient (r), mean absolute error (MAE), and root mean square error (RMSE) variables.

F.V.	G.L.	r	MAE	RMSE
ML	5	0.814 *	0.000 *	0.0004 *
input	3	1.002 *	0.000 *	0.0001 *
ML * input	15	0.133 *	0.000 *	0.0000 *
Waste	216	0.007	0.000	0.0000
Total	239	0.045	0.000	0.0000
C.V. (%)		36.4	11.79	10.28

F.V.: Sources of variation; G.L.: degree of freedom; ML: machine learning algorithms; C.V. (%): coefficient of variation. *: significant at 5% de probability.

The ML models used spectral data from coffee beans (CG) and spectral data from these same samples after grinding and roasting (CGRG) and were submitted to machine learning models for caffeine content prediction. In addition to these inputs, the specification of clones (CG+C and CGRG+C) was also tested, totaling four different inputs (Figure 5). In terms of Pearson's correlation coefficient, which indicates prediction accuracy, the inputs tested with the additional clone information, along with the spectral data provided the best accuracy for DT and M5P, reaching an average of 0.45 (Figure 6). RF and ANN models performed better when using the spectral information of the coffee beans along with the clone, with accuracies close to 0.4. SVM showed the best accuracy when using spectral information from the ground and roasted beans to identify the clones, with r-values close to 0.7, showing the best accuracy for caffeine content prediction.

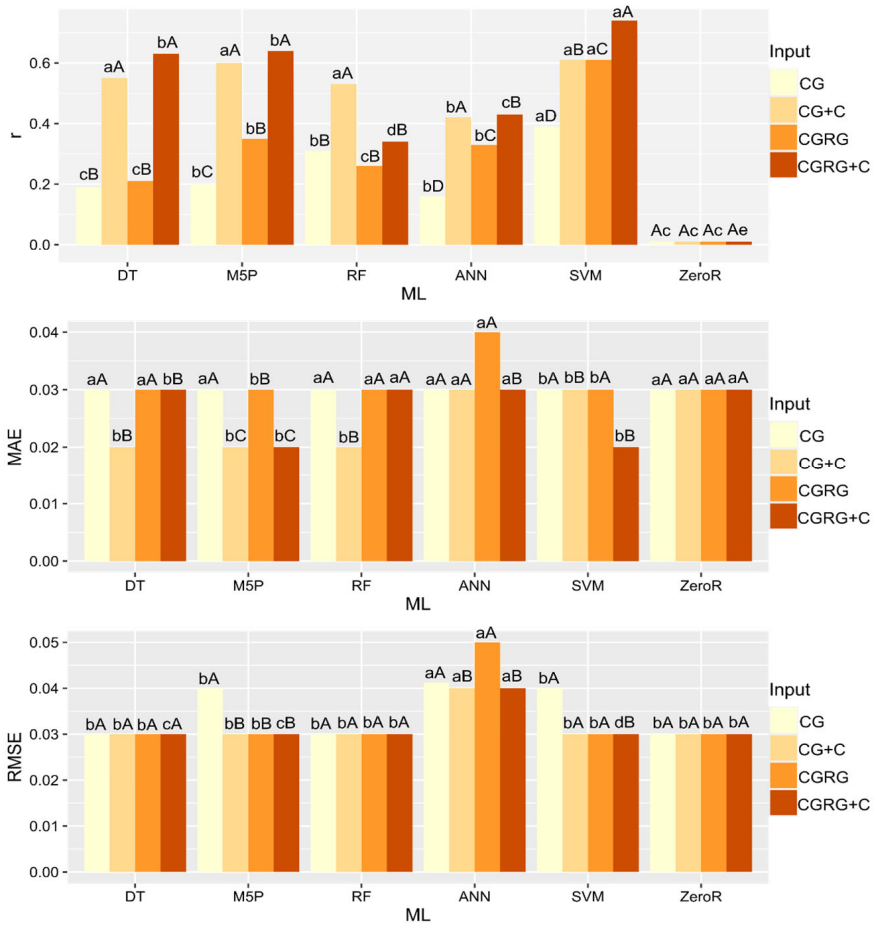

Figure 5. Bar graphs for comparing the means of the precision measures correlation coefficient (r), mean absolute error (MAE), and root mean square error (RMSE) for predicting caffeine content in coffee beans. Spectral information of the grain (CG), spectral information of the grain with additional clone information (CG+C), spectral information of the grain after roasting and grinding (CGRG), and spectral information of the grain after roasting and grinding with additional clone information (CGRG+C). Equal uppercase letters do not differ among the inputs tested; equal lowercase letters do not differ among the machine learning models. Mean comparisons were made using the Scott-Knot grouping at 5% probability.

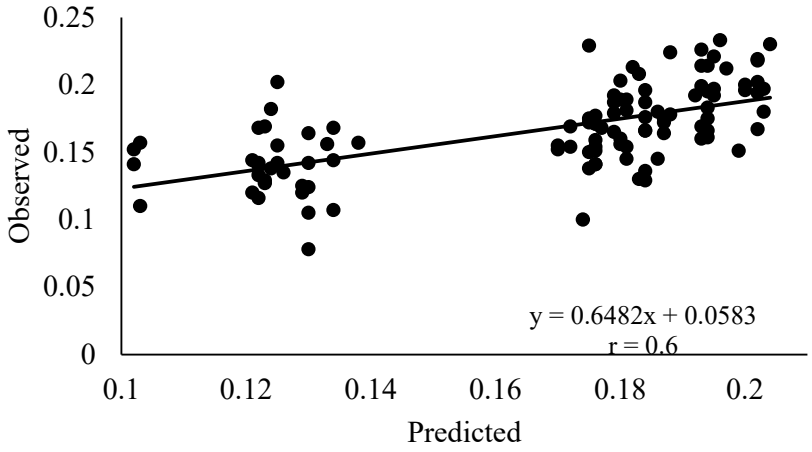

Figure 6. Observed versus predicted values for caffeine content prediction obtained by the best model and best input.

For the behavior of the inputs for each model, the use of CG provided the best accuracy for SVM, with a value close to 0.20 being the best accuracy among the models but with low prediction results. When this same information was used as an input with additional clone identification, the DT, M5P, RF and SVM models showed good accuracy, with performances above 0.40. When the input used was the spectral information of the grains after roasting and grinding, the best model was SVM, with an accuracy close to 0.5. The CGRG+C SVM information also achieved the best accuracy, with values of approximately 0.7.

Comparing the four inputs to the models, it can be seen that, for DT and RF, CG+C provided the lowest error rate for the algorithms. M5P performed well in terms of error when clone identification was added to the inputs. SVM showed a lower error with CGRG+C. Comparing each input among the models in terms of MAE error, SVM showed lower errors (0.02) when using CG spectral information. ZeroR showed the same behavior across all the models. Using coffee bean spectral information with clone identification provided the lowest error for most algorithms: SVM, DT, M5P and RF, with an average error of 0.02 for all of them. Using the spectral information of the beans after grinding and roasting, the M5P and SVM models had the lowest errors (0.02). Adding clone information to this last input, in addition to M5P and SVM, the DT model also showed lower MAE values (0.02).

In terms of RMSE, all the inputs tested had the same error rate for DT, RF and ZeroR. M5P and SVM showed a lower error rate when using CG+C, CGRG and CGRG+C. ANN showed the lowest error rate with CG, CG+C and CGRG+C. Evaluating the error performance that each input generated for the models, the use of all inputs was lower for all models except ANN.

The algorithm with the best performance was SVM using CGRG+C as input, so a regression graph was constructed to express predicted and observed data using the algorithm and input with the best performance (Figure 6). The model's prediction accuracy was 0.6 for the correlation coefficient and 0.02 for the mean absolute error and root mean square error.

3.3. Comparison of Different Methods

Once the best input for caffeine content prediction had been obtained, it was resubmitted to the ML models with different sample sizes: all the samples evaluated (ALL), half the samples evaluated (50%) and two-thirds of the samples evaluated (75%). Pearson's correlation coefficient shows that regardless of the number of samples submitted to the models, they have the same accuracy behavior, except for ANN and SVM, which demonstrated better accuracy with the larger amount of data used (Figure 7). Evaluating the individual behavior of each model with each input reveals that using ALL, SVM achieved the best

prediction performance and when the database was halved, DT, M5P and SVM performed well; the same algorithms exhibited good accuracy when 75% of the database was used.

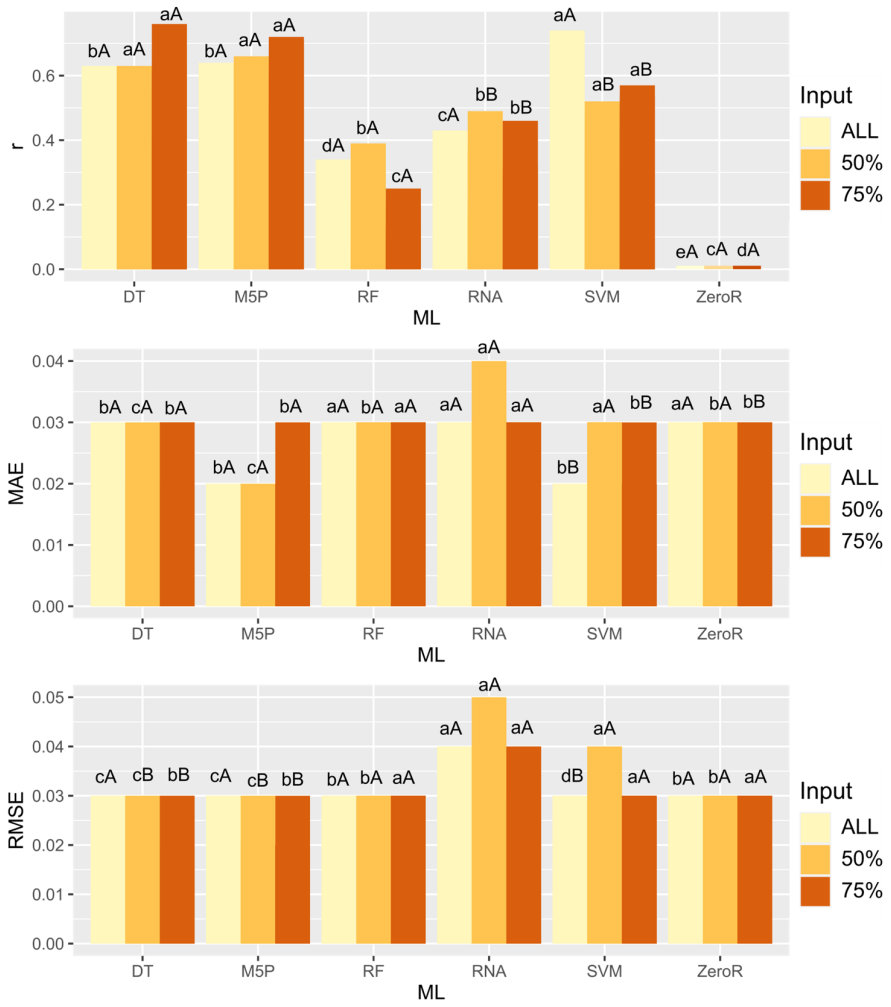

Figure 7. Bar graphs for comparing the means of the precision measures correlation coefficient (r), mean absolute error (MAE) and root mean square error (RMSE) for the prediction of caffeine content in coffee beans in different numbers of samples. Equal uppercase letters do not differ among the inputs tested; equal lowercase letters do not differ among the machine learning models. Mean comparisons were made using the Scott-Knot grouping at 5% probability.

As for the error accuracy metric, for MAE comparing the three inputs and for most of the models, there was no difference among the number of samples, except for SVM, which exhibited a lower error rate with larger numbers of samples. The use of ALL provided the lowest error rate for DT, M5P and SVM; the use of 50% of the database provided the lowest error rate for DT and M5P and both algorithms had low error when 75% of the database was used. For RMSE, using 50% and 75% of the complete database provided lower errors for DT and M5P; ALL provided lower errors for SVM; the other models showed similar behavior regardless of the input used. Comparing the performance of the models with each input, it is important to note that SVM exhibited the lowest RMSE when ALL was used; when 50% was used as input, the lowest errors were shown by DT and M5P, as well as when 75% of the database was used as input. In summary, SVM performed best with all samples, while DT and M5P were more effective with 50% and 75%.

4. Discussion

Caffeine plays a significant role in the human food chain through plant-derived foods, such as tea, coffee beans, cocoa beans and nuts. Caffeine is an alkaloid found mainly in coffee and is used abundantly as an additive in foods, beverages and medicines [19]. Differences in caffeine content are common among coffee samples of different varieties, seasons and roasting or grinding processes [20]. *Coffea arabica* is the most cultivated and consumed species, specifically in Brazil, well known for its fine and exquisite aromas, while *Coffea canephora* has specific characteristics, such as a higher caffeine content and distinct flavors [3,21,22]. This result is demonstrated in Figure 3, which shows a variation in caffeine content among the clones studied, due to the factors mentioned above.

The chemical composition of coffee varies widely between Arabica (*Coffea arabica*) and Robusta (*Coffea canephora*), which can be influenced by environmental factors, geographical origin and post-harvest processing, not to mention the notable differences within the same batch [1,23]. This variability between *Coffea arabica* and *Coffea canephora* is important so that the algorithms have data with greater variability and can learn more accurately and effectively to predict caffeine content.

The use of bean reflectance allows coffee beans to be analyzed in nature, preserving the samples for other analyses or marketing, since the use of chromatography can be time-consuming and costly. The use of the hyperspectral sensor for this function can be applied at different stages of the coffee production chain, from the selection of the beans to the quality control of the final product, helping to standardize and improve the quality of the coffee offered to the consumer. Using the reflectance of coffee after it has been roasted and ground helps improve the accuracy of models in predicting caffeine content. This is because coffee beans behave differently and produce different results in terms of physical properties, chemical composition and biological activities when roasted under different conditions [24,25].

After roasting, coffee beans have a higher correlation with caffeine content, indicating that the use of reflectance analysis of roasted and ground beans is more accurate. Roasting profoundly alters the physical and chemical properties of the beans, intensifying the spectral characteristics that are directly linked to caffeine content. Therefore, the reflectance of the beans after roasting and milling proved to be better for increasing the accuracy of the prediction models, providing more reliable estimates. In addition, the identification of clones in the database helped to significantly improve the performance of the algorithms.

The most notable feature of the spectral information forming the hyperspectral curve for each clone is that the entire coffee spectrum shows greater absorbance than reflectance. In addition, the region at approximately 1400 nm is of great importance in predicting caffeine content, chlorogenic acid and total phenolics, and the region at approximately 1200 nm makes important contributions to the model's loadings, especially for the control of chlorogenic acid and total phenolics [11].

The SVM algorithm performed the best in predicting caffeine content using the spectral information mentioned above. The use of ML models is a more advanced approach that efficiently deals with the non-linearity of spectral information with caffeine content and offers robustness against outliers and noise [26]. The SVM is recognized for its robustness, especially against overfitting, an essential feature when dealing with spectral data where the number of features often exceeds the number of observations [27]. In addition to a high r, the model showed low MAE and RMSE values, which indicates that the model is more reliable and consistent in its predictions, providing more precise and accurate results, demonstrating that the predictions generated by the model are closer to the actual observed values and reflecting its effectiveness in generalizing training data to new unseen data [26].

The amount of data affects the performance of ML models differently, depending on the algorithm (Figure 6). SVM's superior performance with the use of all samples (ALL) suggests that this model benefits from larger data sets, supporting what was previously mentioned with the use of this model. DT and M5P maintain good performance with smaller sets (50% and 75%), highlighting the robustness of these models in cases where

there are few samples available. This may be relevant in situations where data collection is difficult, allowing simpler models to still provide reliable results. The accuracy means obtained for the best algorithm and input (SVM using the ALL dataset) suggest that future research should test other experimental conditions, such as larger datasets, different input configurations and different and/or larger amounts of genetic materials in order to increase the accuracy of caffeine content prediction.

The use of hyperspectral data combined with appropriate machine learning models has significant potential and diverse applications in the coffee industry due to its non-destructive nature, rapid analysis capabilities and sophistication in data analysis techniques [26]. The methodology suggested here may have future applications in routine quality control, allowing the rapid assessment of coffee bean quality throughout production, from green beans to roasted coffee, as well as the detection of defects such as mold or insect damage and the determination of other chemical components of coffee.

5. Conclusions

The traditional assessment of caffeine content determination requires laboratory procedures that are often expensive, time-consuming and specialized infrastructure. These limitations make the process more complex for determination, especially on a large scale or for studies that require a high sampling frequency. The proposed method, using machine learning models with spectral data, offers an efficient and low-cost long-term solution. The combination of both technologies makes it possible to predict caffeine content quickly and with good accuracy, even with small data sets. By applying the right model, this approach can optimize both time and financial resources. The support vector machine (SVM) algorithm demonstrated the best accuracy in predicting caffeine content by using hyperspectral data from roasted and ground coffee beans. This performance was improved significantly when the clones were identified, with values of approximately 0.6 for the correlation coefficient and with a low mean absolute error of approximately 0.02.

Author Contributions: Conceptualization, D.C.S. and R.F.R.; methodology, F.L.Z., C.A.d.S.J. and P.E.T.; software, D.C.S., A.C.d.S.C.S., C.N.S.C. and M.B.; validation, P.E.T., R.G.d.S., C.N.S.C. and R.d.C.F.A.; formal analysis, D.C.S. and C.N.S.C.; investigation, N.P.d.S. and G.S.O.; resources, P.E.T. and A.C.d.S.C.S.; data curation, P.E.T.; writing—original draft preparation, D.C.S.; writing—review and editing, C.N.S.C.; visualization, L.P.R.T. and P.E.T.; supervision, R.F.R. and P.E.T.; project administration, P.E.T. and A.C.d.S.C.S.; funding acquisition, C.A.d.S.J. and P.E.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

ML	machine learning
CG	Spectral information of the bean
CG+C	Spectral information of the bean with additional clone information (CG+C)
CGRG	spectral information of the bean after roasting and grinding
CGRG+C	spectral information of the bean after roasting and grinding with additional clone information
ANN	artificial neural networks
DT	decision tree
LR	linear regression
RF	random forest
SCA	Specialty Coffee Association

UPLC	ultra-performance liquid chromatography
F.V.	Sources of variation
G.L.	degree of freedom
C.V.	coefficient of variation
r	Pearson correlation coefficient
MAE	mean absolute error
RMSE	root mean square error

References

1. Caporaso, N.; Whitworth, M.B.; Grebby, S.; Fisk, I.D. Non-Destructive Analysis of Sucrose, Caffeine and Trigonelline on Single Green Coffee Beans by Hyperspectral Imaging. *Food Res. Int.* **2018**, *106*, 193–203. [CrossRef] [PubMed]
2. Eron, F.; Noman, M.; de Oliveira, R.R.; Chalfun-Junior, A. Computer Vision-Aided Intelligent Monitoring of Coffee: Towards Sustainable Coffee Production. *Sci. Hortic.* **2024**, *327*, 112847. [CrossRef]
3. Freitas, V.V.; Borges, L.L.R.; Vidigal, M.C.T.R.; dos Santos, M.H.; Stringheta, P.C. Coffee: A Comprehensive Overview of Origin, Market, and the Quality Process. *Trends Food Sci. Technol.* **2024**, *146*, 104411. [CrossRef]
4. Sualeh, A.; Tolessa, K.; Mohammed, A. Biochemical Composition of Green and Roasted Coffee Beans and Their Association with Coffee Quality from Different Districts of Southwest Ethiopia. *Heliyon* **2020**, *6*, e05812. [CrossRef]
5. Loukri, A.; Sarafera, C.; Goula, A.M.; Gardikis, K.; Mourtzinos, I. Green Extraction of Caffeine from Coffee Pulp Using a Deep Eutectic Solvent (DES). *Appl. Food Res.* **2022**, *2*, 100176. [CrossRef]
6. Ayu, P.C.; Budiastra, I.W.; Rindang, A. NIR Spectroscopy Application for Determination Caffeine Content of Arabica Green Bean Coffee. In *IOP Conference Series: Earth and Environmental Science*; IOP Publishing: Bristol, UK, 2020; Volume 454, p. 012049.
7. Miras-Moreno, B.; Monterisi, S.; Rouphael, Y.; Colla, G.; Lucini, L.; Cesco, S.; Pii, Y. Integrated Metabolomics and Morpho-Biochemical Analyses Reveal a Better Performance of *Azospirillum brasilense* over Plant-Derived Biostimulants in Counteracting Salt Stress in Tomato. *Int. J. Mol. Sci.* **2022**, *23*, 14216. [CrossRef]
8. Arai, K.; Terashima, H.; Aizawa, S.; Taga, A.; Yamamoto, A.; Tsutsumiuchi, K.; Kodama, S. Simultaneous Determination of Trigonelline, Caffeine, Chlorogenic Acid and Their Related Compounds in Instant Coffee Samples by HPLC Using an Acidic Mobile Phase Containing Octanesulfonate. *Anal. Sci.* **2015**, *31*, 831–835. [CrossRef]
9. Craig, A.P.; Fields, C.; Liang, N.; Kitts, D.; Erickson, A. Performance Review of a Fast HPLC-UV Method for the Quantification of Chlorogenic Acids in Green Coffee Bean Extracts. *Talanta* **2016**, *154*, 481–485. [CrossRef] [PubMed]
10. Fajara, B.E.P.; Susanti, H. HPLC Determination of Caffeine in Coffee Beverage. In Proceedings of the IOP Conference Series: Materials Science and Engineering; IOP Publishing: Bristol, UK, 2017; Volume 259, p. 012011.
11. Nogales-Bueno, J.; Baca-Bocanegra, B.; Romero-Molina, L.; Martínez-López, A.; Rato, A.E.; Heredia, F.J.; Hernández-Hierro, J.M.; Escudero-Gilete, M.L.; González-Miret, M.L. Control of the Extractable Content of Bioactive Compounds in Coffee Beans by near Infrared Hyperspectral Imaging. *LWT* **2020**, *134*, 110201. [CrossRef]
12. Caporaso, N.; Whitworth, M.B.; Fisk, I.D. Prediction of Coffee Aroma from Single Roasted Coffee Beans by Hyperspectral Imaging. *Food Chem.* **2022**, *371*, 131159. [CrossRef]
13. Syed, T.A.; Ansari, K.B.; Banerjee, A.; Wood, D.A.; Khan, M.S.; Al Mesfer, M.K. Machine-learning Predictions of Caffeine Co-crystal Formation Accompanying Experimental and Molecular Validations. *J. Food Process Eng.* **2023**, *46*, e14230. [CrossRef]
14. Beitollahi, M.; Hosseini, S.A. Using Savitsky-Golay Smoothing Filter in Hyperspectral Data Compression by Curve Fitting. In Proceedings of the Iranian Conference on Electrical Engineering (ICEE), Mashhad, Iran, 8–10 May 2018; pp. 452–457.
15. De Gregori, G.S.; de Souza Loureiro, E.; Amorim Pessoa, L.G.; de Azevedo, G.B.; Azevedo, G.T.d.O.S.; Santana, D.C.; de Oliveira, I.C.; de Oliveira, J.L.G.; Teodoro, L.P.R.; Baio, F.H.R. Machine Learning in the Hyperspectral Classification of *Glycaspis brimblecombei* (Hemiptera Psyllidae) Attack Severity in Eucalyptus. *Remote Sens.* **2023**, *15*, 5657. [CrossRef]
16. Santana, D.C.; dos Santos, R.G.; da Silva, P.H.N.; Pistori, H.; Teodoro, L.P.R.; Poersch, N.L.; de Azevedo, G.B.; de Oliveira Sousa Azevedo, G.T.; da Silva Junior, C.A.; Teodoro, P.E. Machine Learning Methods for Woody Volume Prediction in Eucalyptus. *Sustainability* **2023**, *15*, 10968. [CrossRef]
17. Pereira Ribeiro Teodoro, L.; Estevão, R.; Santana, D.C.; de Oliveira, I.C.; Lopes, M.T.G.; de Azevedo, G.B.; Rojo Baio, F.H.; da Silva Junior, C.A.; Teodoro, P.E. Eucalyptus Species Discrimination Using Hyperspectral Sensor Data and Machine Learning. *Forests* **2023**, *15*, 39. [CrossRef]
18. Bouckaert, R.R.; Frank, E.; Hall, M.A.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. WEKA—Experiences with a Java Open-Source Project. *J. Mach. Learn. Res.* **2010**, *11*, 2533–2541.
19. Zareef, M.; Hassan, M.M.; Arslan, M.; Ahmad, W.; Ali, S.; Ouyang, Q.; Li, H.; Wu, X.; Chen, Q. Rapid Prediction of Caffeine in Tea Based on Surface-Enhanced Raman Spectroscopy Coupled Multivariate Calibration. *Microchem. J.* **2020**, *159*, 105431. [CrossRef]
20. Mori, A.L.B.; Viegas, M.C.; Ferrão, M.A.G.; Fonseca, A.F.; Ferrão, R.G.; Benassi, M.T. Coffee Brews Composition from *Coffea canephora* Cultivars with Different Fruit-Ripening Seasons. *Br. Food J.* **2020**, *122*, 827–840. [CrossRef]
21. Poisson, L.; Blank, I.; Dunkel, A.; Hofmann, T. The Chemistry of Roasting—Decoding Flavor Formation. In *The Craft and Science of Coffee*; Elsevier: Amsterdam, The Netherlands, 2017; pp. 273–309.
22. Hall, R.D.; Trevisan, F.; de Vos, R.C.H. Coffee Berry and Green Bean Chemistry–Opportunities for Improving Cup Quality and Crop Circularity. *Food Res. Int.* **2022**, *151*, 110825. [CrossRef]

23. Joët, T.; Laffargue, A.; Descroix, F.; Doulbeau, S.; Bertrand, B.; Dussert, S. Influence of Environmental Factors, Wet Processing and Their Interactions on the Biochemical Composition of Green Arabica Coffee Beans. *Food Chem.* **2010**, *118*, 693–701. [CrossRef]
24. Baggenstoss, J.; Poisson, L.; Kaegi, R.; Perren, R.; Escher, F. Coffee Roasting and Aroma Formation: Application of Different Time—Temperature Conditions. *J. Agric. Food Chem.* **2008**, *56*, 5836–5846. [CrossRef]
25. da Rosa, J.S.; Freitas-Silva, O.; Rouws, J.R.C.; da Silva Moreira, I.G.; Novaes, F.J.M.; de Almeida Azevedo, D.; Schwab, N.; de Oliveira Godoy, R.L.; Eberlin, M.N.; de Rezende, C.M. Mass Spectrometry Screening of Arabica Coffee Roasting: A Non-Target and Non-Volatile Approach by EASI-MS and ESI-MS. *Food Res. Int.* **2016**, *89*, 967–975. [CrossRef]
26. Munawar, A.A.; Mörlein, D. Prediction Accuracy of Near Infrared Spectroscopy Coupled with Adaptive Machine Learning Methods for Simultaneous Determination of Chlorogenic Acid and Caffeine on Intact Coffee Beans. *Case Stud. Chem. Environ. Eng.* **2024**, *10*, 100913. [CrossRef]
27. Xu, S.; Lu, B.; Baldea, M.; Edgar, T.F.; Nixon, M. An Improved Variable Selection Method for Support Vector Regression in NIR Spectral Modeling. *J. Process Control* **2018**, *67*, 83–93. [CrossRef]